

# Shubham Agarwal

✉ shubham3@berkeley.edu    Github    Google Scholar    LinkedIn    Website

## RESEARCH INTEREST

My research lies at the intersection of **Machine Learning and Systems**, focusing on LLMs and agentic systems. I develop **structured memory** and **evolutionary approaches** to improve agent learning and reliability, and work on **long-context inference** and attention analysis in LLMs to make agents more efficient and trustworthy.

## EDUCATION

**Ph.D. in Computer Science** 2025 – present  
University of California, Berkeley - Advisor: *Prof. Ion Stoica, Prof. Aditya Parameswaran*

**Bachelors in Computer Science** 2018 – 2022  
Birla Institute of Technology and Science (BITS) Pilani, Pilani Campus, India GPA: 9.95/10.00

## RESEARCH EXPERIENCE

**Research Associate II** – Adobe Research, India Jul 2022 – Aug 2025  
Systems and Insights Group. Mentors: *Dr. Subrata Mitra, Dr. Shiv K Saini*

- Designed a high throughput inference serving system for generative models by employing caching techniques.
- Developed an outage forecasting and root cause diagnosis pipeline for microservice-based cloud systems.
- Published **7 papers**, filed **6 patents**, and successfully integrated research innovations into **3 products**.

**Research Intern** – American Express, India Jan 2022 – Jul 2022  
Document Analytics & Intelligence Lab. Mentor: *Dr. Himanshu S Bhatt*

- Designed a multi-modal deep learning pipeline for processing visually rich documents by customizing LayoutLM.
- Developed a few-shot learning pipeline for rapid adaptation to new document types with synthetic data augmentation.
- Implemented a collaborative inference framework using server-side PyTorch and on-device TFLite models.

**Research Intern** – Adobe Research, India May 2021 – Aug 2021  
Big Data Intelligence Lab. Mentor: *Dr. Shiv K Saini*

- Modeled the degradation in QoS metrics distribution as a function of system metrics to forecast rare outages.
- Used a Mixture Density Network trained with tail regularization to improve precision in distribution forecasting.
- Resulted in a full research paper **publication** at FSE 2023 and a US **patent** issuance.

## PUBLICATIONS

\* equal contribution

- [1] **S Agarwal\***, S Sundaresan\*, S Mitra, D Mahapatra, A Gupta, R Sharma, NJ Kapu, T Yu, S Saini. "Cache-craft: Managing chunk-caches for efficient retrieval-augmented generation." In *The Proceedings of the ACM on Management of Data*. [SIGMOD 2025](#). (acceptance rate 20%)
- [2] A Ikram\*, K Lee\*, **S Agarwal**, S Saini, S Bagchi, and M Kocaoglu. "Root Cause Analysis of Failures from Partial Causal Structures." In *The 41st Conference on Uncertainty in Artificial Intelligence*. [UAI 2025](#). (acceptance rate 30%)
- [3] **S Agarwal**, S Mitra, S Chakraborty, S Karanam, K Mukherjee, and S Saini. "Approximate Caching for Efficiently Serving Text-to-Image Diffusion Models." In *The 21st USENIX Symposium on Networked Systems Design and Implementation*. [NSDI 2024](#). (acceptance rate 18%)
- [4] C Lu\*, **S Agarwal\***, M Tanjim, K Mahadik, A Rao, S Mitra, S Saini, S Bagchi, S Chatterji. "ReCon: Training-Free Acceleration for Text-to-Image Synthesis with Retrieval of Concept Prompt Trajectories." In *The 18th European Conference on Computer Vision*. [ECCV 2024](#). (acceptance rate 27.9%)
- [5] G S Ahmad, **S Agarwal**, S Mitra, R A Rossi, M Doshi, and S Paila. "ScaleViz: Scaling Visualization Recommendation Models on Large Data." In *The 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [PAKDD 2024](#). (Oral acceptance 18.75%)
- [6] **S Agarwal**, S Chakraborty, S Garg, S Bisht, C Jain, A Gonuguntla, and S Saini. "Outage-Watch: Early Prediction of Outages using Extreme Event Regularizer." In *The 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. [FSE 2023](#). (acceptance rate 25.6%)
- [7] S Chakraborty, **S Agarwal**, S Garg, A Sethia, U Pandey, V Aggarwal, and S Saini. "ESRO: Experience Assisted Service Reliability against Outages." In *The 38th IEEE/ACM International Conference on Automated Software Engineering*. [ASE 2023](#). (acceptance 21.3%)
- [8] S Chakraborty, S Garg\*, **S Agarwal\***, A Chauhan, and S Saini. "CausIL: Causal Graph for Instance Level Microservice Data." In *Proceedings of The Web Conference*. [WWW 2023](#). (acceptance rate 19.2%)
- [9] **S Agarwal**, G Chan, S Garg, T Yu, and S Mitra. "Fast Natural Language Based Data Exploration with Samples." In *Companion of the 2023 International Conference on Management of Data*. [SIGMOD 2023](#) (Demo track).

Under review:

- [1] **S Agarwal**, S Iqbal, and S Mitra. "Micro-managing Prompts for High-Throughput Text-to-Image Inference Serving using Approximation." (*accepted with shepherding at Middleware'25*)
- [2] D Mahapatra\*, **S Agarwal\***, A Saxena, S Mitra. "RCStat: A Statistical Framework for using Relative Contextualization in Transformers." (*submitted to ICLR'25*)

---

## PATENTS

- [1] D Mahapatra, **S Agarwal**, et al., Relative Contextualization using Pre-Softmax Logits in Transformers. [US Patent 19/285,322]
- [2] **S Agarwal**, et al., Managing Chunk Caches for Efficient Retrieval-Augmented Generation in LLMs. [US Patent 19/074,061]
- [3] C Lu, **S Agarwal**, et al., Recon: Acceleration for Text-to-Image Synthesis with Concept Trajectories. [US Patent 63/698,535]
- [4] **S Agarwal**, S Mitra, S Iqbal, Micromanaging Prompts for High-Throughput Text-to-Image Inference. [US Patent 18/808,654]
- [5] **S Agarwal**, S Mitra, S Chakraborty, et al., Intermediate Noise Retrieval for Image Generation. [US Patent 18/637,024]
- [6] **S Agarwal**, S Mitra, et al., Using Reinforcement Learning to recommend Data Visualizations. [US Patent 18/668,888]
- [7] S Mitra, **S Agarwal**, G Chan, et al., Data Exploration using Natural Language with Data Sampling. [US Patent 18/675,930]
- [8] S Garg, **S Agarwal**, S Bisht, N Sheoran, et al., A System and Method for Outage Forecasting. [US Patent 17/656,263]

---

## SELECTED PROJECTS

### Efficient Inference Strategies (Systems for ML)

*Papers published in [NSDI '24, ECCV '24, PAKDD '24]*

- **Efficient Chunk-Caching for LLM-based RAG**: Developed a Key-Value reuse mechanism that decomposes prefill states into reusable chunks, achieving a  $2.7\times$  reduction in prefill compute costs and latency while maintaining output quality.
- **Approximate Caching for Diffusion Models (In Product)**: Developed a caching framework for diffusion models, achieving **19.8%** latency reduction. Utilized a concept decomposition technique to improve generation fidelity. Implemented an accuracy scaling framework with query-aware routing, boosting system throughput by **30%** without compromising quality.
- **Scaling VizRec Models on Large Data**: Designed a plug-in framework for Visualization Recommender systems that achieve up to **10 $\times$**  speedup in latency. It implements a scalable Deep Q-Learning to optimize input statistics selection.

### Outage Detection and Diagnosis (ML for Systems)

*Papers published in [FSE '23, ASE '23, WWW '23]*

- **Cloud Outage Prediction (In Product)**: Developed a model to predict QoS metrics distribution, reducing outage detection time by **88%**. The system incorporates real-time metric retrieval from Prometheus to streamline inference and retraining.
- **Root Cause Diagnosis and Resolution**: Developed a novel causal discovery technique and an automated framework linking incidents to historical cases, achieving a **27%** improvement in root cause identification and remediation recommendations.
- **Query Latency Prediction (In Product)**: Built a pipeline to predict Spark job latency, now used by over **50+** engineers for real-time monitoring. The tool ingests job metadata from Kafka and uses Azure functions to scale the deployment.

---

## RESEARCH COLLABORATIONS

- **Training-free acceleration for text-to-image synthesis using prompt trajectories** Oct 2023–Jul 2024  
with **Prof. Somali Chaterji** and **Prof. Saurabh Bagchi**, Purdue University  
- Designed a novel concept retrieval-based method for accelerating diffusion models (*paper at ECCV '24*).
- **Causality-based Root Cause Diagnosis for real-world microservice architectures** Jul 2023–Mar 2025  
with **Prof. Saurabh Bagchi**, CRISP Lab, Purdue University  
- Developed a graph-based causal inference algorithm for RCA in complex systems (*paper at UAI '25*).

---

## TECHNICAL SKILLS

- Programming languages: Python, C, C++, Triton, Java, SQL, Verilog, MIPS
- Packages & Frameworks: PyTorch, Keras, TensorFlow, scikit-learn, Git, Docker, Kubernetes, Kafka

---

## UNDERGRADUATE COURSES:

Operating Systems, Database Systems, Computer Networks, Compilers, Computer Architecture, Artificial Intelligence, Information Retrieval, Data Structures and Algorithms, Probability and Statistics, Linear Algebra

---

## ACTIVITIES & ACHIEVEMENTS

- Industry mentor for UMass Master's students' group, guiding research on RAG systems and on-device model security.
- Mentored over 10 undergraduate interns and collaborated with 3 PhD interns during summer internships at Adobe.
- Contributed to multiple paper reviews for technical conferences including ATC, AISTATS, AAAI, Middleware, etc.
- Co-led cross-institutional collaborations with researchers from Purdue and UMass, organizing workshops and talks.
- Awarded with 100% merit scholarship for academic excellence across all semesters; ranking top among all students.
- Served as the Student Representative for the Student-Faculty Council, representing over 1,000 undergraduate students.
- Successfully organized BITS Pilani's Annual Technical Fest, boosting external participation and outreach by 200%.